Marginality: a numerical mapping for enhanced treatment of nominal and hierarchical attributes [*]

Josep Domingo-Ferrer

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
Tel.: +34 977558270
Fax: +34 977559710
E-mail josep.domingo@urv.cat

Abstract

The purpose of statistical disclosure control (SDC) of microdata, a.k.a. data anonymization or privacy-preserving data mining, is to publish data sets containing the answers of individual respondents in such a way that the respondents corresponding to the released records cannot be re-identified and the released data are analytically useful. SDC methods are either based on masking the original data, generating synthetic versions of them or creating hybrid versions by combining original and synthetic data. The choice of SDC methods for categorical data, especially nominal data, is much smaller than the choice of methods for numerical data. We mitigate this problem by introducing a numerical mapping for hierarchical nominal data which allows computing means, variances and covariances on them.
Keywords: Statistical disclosure control; Data anonymization; Privacy-preserving data mining; Variance of hierarchical data; Hierarchical nominal data

1 Introduction

Statistical disclosure control (SDC, [1, 4, 8, 3, 5]), a.k.a. data anonymization and sometimes as privacy-preserving data mining, aims at making possible the

[*] This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 project "DwB". The author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia.

publication of statistical data in such a way that the individual responses of specific users cannot be inferred from the published data and background knowledge available to intruders. If the data set being published consists of records corresponding to individuals, usual SDC methods operate by masking original data (via perturbation or detail reduction), by generating synthetic (simulated) data preserving some statistical features of the original data or by producing hybrid data obtained as a combination of original and synthetic data. Whatever the protection method chosen, the resulting data should still preserve enough analytical validity for their publication to be useful to potential users.

A microdata set can be defined as a file with a number of records, where each record contains a number of attributes on an individual respondent. Attributes can be classified depending on their range and the operations that can be performed on them:

1. *Numerical.* An attribute is considered numerical if arithmetical operations can be performed on it. Examples are income and age. When designing methods to protect numerical data, one has the advantage that arithmetical operations are possible, and the drawback that every combination of numerical values in the original data set is likely to be unique, which leads to disclosure if no action is taken.

2. *Categorical.* An attribute is considered categorical when it takes values over a finite set and standard arithmetical operations on it do not make sense. Two main types of categorical attributes can be distinguished:

   (a) *Ordinal.* An ordinal attribute takes values in an ordered range of categories. Thus, the $\leq$, max and min operators are meaningful and can be used by SDC techniques for ordinal data. The instruction level and the political preferences (left-right) are examples of ordinal attributes.

   (b) *Nominal.* A nominal attribute takes values in an unordered range of categories. The only possible operator is comparison for equality. Nominal attributes can further be divided into two types:

      i. *Hierarchical.* A hierarchical nominal attribute takes values from a hierarchical classification. For example, plants are classified using Linnaeus's taxonomy, the type of a disease is also selected from a hierarchical taxonomy, and the type of an attribute can be selected from the hierarchical classification we propose in this section.

      ii. *Non-hierarchical.* A non-hierarchical nominal attribute takes values from a flat hierarchy. Examples of such attributes could be the preferred soccer team, the address of an individual, the civil status (married, single, divorced, widow/er), the eye color, etc.

This paper focuses on finding a numerical mapping of nominal attributes, and more precisely hierarchical nominal attributes. In addition to other conceivable applications not dealt with in this paper, such a mapping can be used to anonymize nominal data in ways so far reserved to numerical data. The interest of this is that many more SDC methods exist for anonymizing numerical data than categorical and especially nominal data.

Assuming a hierarchy is less restrictive than it would appear, because very often a non-hierarchical attribute can be turned into a hierarchical one if its flat hierarchy can be developed into a multilevel hierarchy. For instance, the preferred soccer and the address of an individual have been mentioned as non-hierarchical attributes; however, a hierarchy of soccer teams by continent and country could be conceived, and addresses can be hierarchically clustered by neighborhood, city, state, country, etc. Furthermore, well-known approaches to anonynimization, like k-anonymity [7], assume that any attribute can be generalized, *i.e.* that an attribute hierarchy can be defined and values at lower levels of the hierarchy can be replaced by values at higher levels.

1.1 Contribution and plan of this paper

We propose to associate a number to each categorical value of a hierarchical nominal attribute, namely a form of centrality of that category within the attribute's hierarchy. We show how this allows computation of centroids, variances and covariances of hierarchical nominal data.

Section 2 gives background on the variance of hierarchical nominal attributes. Section 3 defines a tree centrality measure called marginality and presents the numerical mapping. Section 4 exploits the numerical mapping to compute means, variances and covariances of hierarchical nominal data. Conclusions are drawn in Section 5.

2 Background

We next recall the variance measure for hierarchical nominal attributes introduced in [2]. To the best of our knowledge, this is the first measure which captures the variability of a sample of values of a hierarchical nominal attribute by taking into account the semantics of the hierarchy. The intuitive idea is that a set of nominal values belonging to categories which are all children of the same parent category in the hierarchy has smaller variance that a set with children from different parent categories.

Algorithm 1 (Nominal variance in [2])

1. *Let the hierarchy of categories of a nominal attribute X be such that b is the maximum number of children that a parent category can have in the hierarchy.*

2. *Given a sample T_X of nominal categories drawn from X, place them in the tree representing the hierarchy of X. Prune the subtrees whose nodes have no associated sample values. If there are repeated sample values, there will*

be several nominal values associated to one or more nodes (categories) in the pruned tree.

3. *Label as follows the edges remaining in the tree from the root node to each of its children:*

    - *If b is odd, consider the following succession of labels $l_0 = (b-1)/2$, $l_1 = (b-1)/2-1$, $l_2 = (b-1)/2+1$, $l_3 = (b-1)/2-2$, $l_4 = (b-1)/2+2$, $\cdots$, $l_{b-2} = 0$, $l_{b-1} = b-1$.*
    - *If b is even, consider the following succession of labels $l_0 = (b-2)/2$, $l_1 = (b-2)/2+1$, $l_2 = (b-2)/2-1$, $l_3 = (b-2)/2+2$, $l_4 = (b-2)/2-2$, $\cdots$, $l_{b-2} = 0$, $l_{b-1} = b-1$.*
    - *Label the edge leading to the child with most categories associated to its descendant subtree as l_0, the edge leading to the child with the second highest number of categories associated to its descendant subtree as l_1, the one leading to the child with the third highest number of categories associated to its descendant subtree as l_2 and, in general, the edge leading to the child with the i-th highest number of categories associated to its descendant subtree as l_{i-1}. Since there are at most b children, the set of labels $\{l_0, \cdots, l_{b-1}\}$ should suffice. Thus an* edge label *can be viewed as a b-ary digit (to the base b).*

4. *Recursively repeat Step 3 taking instead of the root node each of the root's child nodes.*

5. *Assign to values associated to each node in the hierarchy a* node label *consisting of a b-ary number constructed from the edge labels, more specifically as the concatenation of the b-ary digits labeling the edges along the path from the root to the node: the label of the edge starting from the root is the most significant one and the edge label closest to the specific node is the least significant one.*

6. *Let L be the maximal length of the leaf b-ary labels. Append as many l_0 digits as needed in the least significant positions to the shorter labels so that all of them eventually consist of L digits.*

7. *Let $T_X(0)$ be the set of b-ary digits in the least significant positions of the node labels (the "units" positions); let $T_X(1)$ be the set of b-ary digits in the second least significant positions of the node labels (the "tens" positions), and so on, until $T_X(L-1)$ which is the set of digits in the most significant positions of the node labels.*

8. *Compute the variance of the sample as*

$$Var_H(T_X) = Var(T_X(0)) + b^2 \cdot Var(T_X(1)) + \cdots \\ + b^{2(L-1)} \cdot Var(T_X(L-1)) \quad (1)$$

where $Var(\cdot)$ is the usual numerical variance.

In Section 4.2 below we will show that an equivalent measure can be obtained in a simpler and more manageable way.

3 A numerical mapping for nominal hierarchical data

Consider a nominal attribute X taking values from a hierarchical classification. Let T_X be a sample of values of X. Each value $x \in T_X$ can be associated two numerical values:

- The sample frequency of x;
- Some centrality measure of x within the hierarchy of X.

While the frequency depends on the particular sample, centrality measures depend both on the attribute hierarchy and the sample. Known tree centralities attempt to determine the "middle" of a tree [6]. We are rather interested in finding how far from the middle is each node of the tree, that is, how marginal it is. We next propose an algorithm to compute a new measure of the marginality of the values in the sample T_X.

Algorithm 2 (Marginality of nominal values)

1. *Given a sample T_X of nominal categorical values drawn from X, place them in the tree representing the hierarchy of X. There is a one-to-one mapping between the set of tree nodes and the set of categories where X takes values. Prune the subtrees whose nodes have no associated sample values. If there are repeated sample values, there will be several nominal values associated to one or more nodes (categories) in the pruned tree.*

2. *Let L be the depth of the pruned tree. Associate weight 2^{L-1} to edges linking the root of the hierarchy to its immediate descendants (depth 1), weight 2^{L-2} to edges linking the depth 1 descendants to their own descendants (depth 2), and so on, up to weight $2^0 = 1$ to the edges linking descendants at depth $L-1$ with those at depth L. In general, weight 2^{L-i} is assigned to edges linking nodes at depth $i-1$ with those at depth i, for $i = 1$ to L.*

3. *For each nominal value x_j in the sample, its marginality $m(x_j)$ is defined and computed as*

$$m(x_j) = \sum_{x_l \in T_X - \{x_j\}} d(x_j, x_l)$$

where $d(x_j, x_l)$ is the sum of the edge weights along the shortest path from the tree node corresponding to x_j and the tree node corresponding to x_l.

Clearly, the greater $m(x_j)$, the more marginal (*i.e.* the less central) is x_j. Some properties follow which illustrate the rationale of the distance and the weights used to compute the marginality.

Lemma 1 *$d(\cdot,\cdot)$ is a distance in the mathematical sense.*

Being the length of a path, it is immediate to check that $d(\cdot,\cdot)$ satisfies reflexivity, symmetry and subadditivity. The rationale of the above exponential weight scheme is to give more weight to differences at higher levels of the hierarchy; specifically, the following property is satisfied.

Lemma 2 *The distance between any non-root node n_j and its immediate ancestor is greater than the distance between n_j and any of its descendants.*

Proof: Let L be the depth of the overall tree and L_j be the depth of n_j. The distance between n_j and its immediate ancestor is 2^{L-L_j}. The distance between n_j and its most distant ancestor is

$$1 + 2 + \cdots + 2^{L-L_j-1} = 2^{L-L_j} - 1$$

□

Lemma 3 *The distance between any two nodes at the same depth is greater than the longest distance within the subtree rooted at each node.*

Proof: Let L be the depth of the overall tree and L_j be the depth of the two nodes. The shortest distance between both nodes occurs when they have the same parent and it is

$$2 \cdot 2^{L-L_j} = 2^{L-L_j+1}.$$

The longest distance within any of the two subtrees rooted at the two nodes at depth L_j is the length of the path between two leaves at depth L, which is

$$2 \cdot (1 + 2 + \cdots + 2^{L-L_j-1}) = 2(2^{L-L_j} - 1) = 2^{L-L_j+1} - 2$$

□

4 Statistical analysis of numerically mapped nominal data

In the previous section we have shown how a nominal value x_j can be associated a marginality measure $m(x_j)$. In this section, we show how this numerical magnitude can be used in statistical analysis.

4.1 Mean

The mean of a sample of nominal values cannot be computed in the standard sense. However, it can be reasonably approximated by the least marginal value, that is, by the most central value in terms of the hierarchy.

Definition 1 (Marginality-based approximated mean) *Given a sample T_X of a hierarchical nominal attribute X, the marginality-based approximated mean is defined as*

$$Mean_M(T_X) = \arg \min_{x_j \in T_X} m(x_j)$$

if one wants the mean to be a nominal value, or

$$Num\_mean_M(T_X) = \min_{x_j \in T_X} m(x_j)$$

if one wants a numerical mean value.

4.2 Variance

In Section 2 above, we recalled a measure of variance of a hierarchical nominal attribute proposed in [2] which takes the semantics of the hierarchy into account. Interestingly, it turns out that the average marginality of a sample is an equivalent way to capture the same notion of variance.

Definition 2 (Marginality-based variance) *Given a sample T_X of n values drawn from a hierarchical nominal attribute X, the marginality-based sample variance is defined as*

$$Var_M(T_X) = \frac{\sum_{x_j \in T_X} m(x_j)}{n}$$

The following lemma is proven in the Appendix.

Lemma 4 *The $Var_M(\cdot)$ measure and the $Var_H(\cdot)$ specified by Algorithm 1 in Section 2 are equivalent.*

4.3 Covariance matrix

It is not difficult to generalize the sample variance introduced in Definition 2 to define the sample covariance of two nominal attributes.

Definition 3 (Marginality-based covariance) *Given a bivariate sample $T_{(X,Y)}$ consisting of n ordered pairs of values $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ drawn from the ordered pair of nominal attributes (X, Y), the marginality-based sample covariance is defined as*

$$Covar_M(T_{(X,Y)}) = \frac{\sum_{j=1}^{n} \sqrt{m(x_j)m(y_j)}}{n}$$

The above definition yields a non-negative covariance whose value is higher when the marginalities of the values taken by X and Y are positively correlated: as the values taken by X become more marginal, so become the values taken by Y.

Given a multivariate data set T containing a sample of d nominal attributes $X^1, \cdots, X^d$, using Definitions 2 and 3 yields a covariance matrix $\mathbf{S} = \{s_{jl}\}$, for $1 \leq j \leq d$ and $1 \leq l \leq d$, where $s_{jj} = Var_M(T_j)$, $s_{jl} = Covar_M(T_{jl})$ for $j \neq l$, T_j is the column of values taken by X^j in T and $T_{jl} = (T_j, T_l)$.

We can use the following distance definition for records with numerical, nominal or hierarchical attributes.

Definition 4 (SSE-distance) *The SSE-distance between two records* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *in a data set with* d *attributes is*

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{12}}{(S^2)^d}} \qquad (2)$$

where $(S^2)^l_{12}$ *is the variance of the* l*-th attribute over the group formed by* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*, and* $(S^2)^l$ *is the variance of the* l*-th attribute over the entire data set.*

We prove in the Appendix the following two theorems stating that the distance above satisfies the properties of a mathematical distance.

Theorem 1 *The SSE-distance on multivariate records consisting of nominal attributes based on the nominal variance computed as per Definition 2 is a distance in the mathematical sense.*

Theorem 2 *The SSE-distance on multivariate records consisting of ordinal or numerical attributes based on the usual numerical variance is a distance in the mathematical sense.*

By combining the proofs of Theorems 1 and 2, the next corollary follows.

Corollary 1 *The SSE-distance on multivariate records consisting of attributes of any type, where the nominal variance is used for nominal attributes and the usual numerical variance is used for ordinal and numerical attributes, is a distance in the mathematical sense.*

5 Conclusions

We have presented a centrality-based mapping of hierarchical nominal data to numbers. We have shown how such a numerical mapping allows computing means, variances and covariances of nominal attributes, and distances between records containing any kind of attributes. Such enhanced flexility of manipulation of nominal attributes can be used, *e.g.* to adapt anonymization methods intented for numerical data to the treament of nominal and hierarchical attributes. The only requirement is that, whatever the treatment, it should not modify the numerical values assigned by marginality, in order for the numerical mapping to be reversible and allow recovering the original nominal values after treatment.

Appendix

Proof (Lemma 4): We will show that, given two samples $T_X = \{x_1, \cdots, x_n\}$ and $T'_X = \{x'_1, \cdots, x'_n\}$ of a nominal attribute X, both with the same cardinality n, it holds that $Var_M(T_X) < Var_M(T'_X)$ if and only if $Var_H(T_X) < Var_H(T'_X)$.

Assume that $Var_M(T_X) < Var_M(T'_X)$. Since both samples have the same cardinality, this is equivalent to

$$\sum_{j=1}^{n} m(x_j) < \sum_{j=1}^{n} m(x'_j)$$

By developing the marginalities, we obtain

$$\sum_{j=1}^{n} \sum_{x_l \in T_X - \{x_j\}} d(x_j, x_l) < \sum_{j=1}^{n} \sum_{x'_l \in T'_X - \{x'_j\}} d(x'_j, x'_l)$$

Since distances are sums of powers of 2, from 1 to 2^{L-1}, we can write the above inequality as

$$d_0 + 2d_1 + \cdots + 2^{L-1} d_{L-1} < d'_0 + 2d'_1 + \cdots + 2^{L-1} d'_{L-1} \tag{3}$$

By viewing $d_{L-1} \cdots d_1 d_0$ and $d'_{L-1} \cdots d'_1 d'_0$ as binary numbers, it is easy to see that Inequality (3) implies that some i must exist such that $d_i < d'_i$ and $d_{\hat{i}} \le d'_{\hat{i}}$ for $i < \hat{i} \le L-1$. This implies that there are less high-level edge differences associated to the values of T_X than to the values of T'_X. Hence, in terms of $Var_H(\cdot)$, we have that $Var(T_X(i)) < Var(T'_X(i))$ and $Var(T_X(\hat{i})) \le Var(T'_X(\hat{i})$ for $i < \hat{i} \le L-1$. This yields $Var_H(T_X) < Var_H(T'_X)$.

If we now assume $Var_H(T_X) < Var_H(T'_X)$ we can prove $Var_M(T_X) < Var_M(T'_X)$ by reversing the above argument. □.

Lemma 5 *Given non-negative A, A', A'', B, B', B'' such that $\sqrt{A} \le \sqrt{A'} + \sqrt{A''}$ and $\sqrt{B} \le \sqrt{B'} + \sqrt{B''}$ it holds that*

$$\sqrt{A+B} \le \sqrt{A'+B'} + \sqrt{A''+B''} \tag{4}$$

Proof (Lemma 5): Squaring the two inequalities in the lemma assumption, we obtain

$$A \le (\sqrt{A'} + \sqrt{A''})^2$$
$$B \le (\sqrt{B'} + \sqrt{B''})^2$$

Adding both expressions above, we get the square of the left-hand side of Expression (4)

$$A + B \le (\sqrt{A'} + \sqrt{A''})^2 + (\sqrt{B'} + \sqrt{B''})^2$$

$$= A' + A'' + B' + B'' + 2(\sqrt{A'A''} + \sqrt{B'B''}) \tag{5}$$

Squaring the right-hand side of Expression (4), we get

$$(\sqrt{A' + B'} + \sqrt{A'' + B''})^2$$

$$= A' + B' + A'' + B'' + 2\sqrt{(A' + B')(A'' + B'')} \tag{6}$$

Since Expressions (5) and (6) both contain the terms $A' + B' + A'' + B''$, we can neglect them. Proving Inequality (4) is equivalent to proving

$$\sqrt{A'A''} + \sqrt{B'B''} \leq \sqrt{(A' + B')(A'' + B'')}$$

Suppose the opposite, that is,

$$\sqrt{A'A''} + \sqrt{B'B''} > \sqrt{(A' + B')(A'' + B'')} \tag{7}$$

Square both sides:

$$A'A'' + B'B'' + 2\sqrt{A'A''B'B''} >$$

$$(A' + B')(A'' + B'') = A'A'' + B'B'' + A'B'' + B'A''$$

Subtract $A'A'' + B'B''$ from both sides to obtain

$$2\sqrt{A'A''B'B''} > A'B'' + B'A''$$

which can be rewritten as

$$(\sqrt{A'B''} - \sqrt{B'A''})^2 < 0$$

Since a real square cannot be negative, the assumption in Expression (7) is false and the lemma follows. □

Proof (Theorem 1): We must prove that the SSE-distance is non-negative, reflexive, symmetrical and subadditive (*i.e.* it satisfies the triangle inequality).

Non-negativity. The SSE-distance is defined as a non-negative square root, hence it cannot be negative.

Reflexivity. If $\mathbf{x}_1 = \mathbf{x}_2$, then $\delta(\mathbf{x}_1, \mathbf{x}_2) = 0$. Conversely, if $\delta(\mathbf{x}_2, \mathbf{x}_2) = 0$, the variances are all zero, hence $\mathbf{x}_1 = \mathbf{x}_2$.

Symmetry. It follows from the definition of the SSE-distance.

Subadditivity. Given three records $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, we must check whether

$$\delta(\mathbf{x}_1, \mathbf{x}_3) \overset{?}{\leq} \delta(\mathbf{x}_1, \mathbf{x}_2) + \delta(\mathbf{x}_2, \mathbf{x}_3)$$

By expanding the above expression using Expression (2), we obtain

$$\sqrt{\frac{(S^2)^1_{13}}{(S^2)^1} + \cdots + \frac{(S^2)^d_{13}}{(S^2)^d}} \overset{?}{\leq}$$

$$\sqrt{\frac{(S^2)^1_{12}}{(S^2)^1}+\cdots+\frac{(S^2)^d_{12}}{(S^2)^d}}+\sqrt{\frac{(S^2)^1_{23}}{(S^2)^1}+\cdots+\frac{(S^2)^d_{23}}{(S^2)^d}} \tag{8}$$

Let us start with the case $d = 1$, that is, with a single attribute, *i.e.* $\mathbf{x}_i = x_i$ for $i = 1, 2, 3$. To check Inequality (8) with $d = 1$, we can ignore the variance in the denominators (it is the same on both sides) and we just need to check

$$\sqrt{S^2_{13}} \overset{?}{\leq} \sqrt{S^2_{12}} + \sqrt{S^2_{23}} \tag{9}$$

We have

$$\begin{aligned} S^2_{13} = Var(\{x_1, x_3\}) &= \frac{m(x_1) + m(x_3)}{2} \\ &= \frac{d(x_1, x_3)}{2} + \frac{d(x_3, x_1)}{2} = d(x_1, x_3) \end{aligned} \tag{10}$$

Similarly $S^2_{12} = d(x_1, x_2)$ and $S^2_{23} = d(x_2, x_3)$. Therefore, Expression (9) is equivalent to subadditivity for $d(\cdot, \cdot)$ and the latter holds by Lemma 1. Let us now make the induction hypothesis for $d-1$ and prove subadditivity for any d. Call now

$$A := \frac{(S^2)^1_{13}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{13}}{(S^2)^{d-1}}$$

$$A' := \frac{(S^2)^1_{12}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{12}}{(S^2)^{d-1}}$$

$$A'' := \frac{(S^2)^1_{23}}{(S^2)^1} + \cdots + \frac{(S^2)^{d-1}_{23}}{(S^2)^{d-1}}$$

$$B := \frac{(S^2)^d_{13}}{(S^2)^d};\ B' := \frac{(S^2)^d_{12}}{(S^2)^d};\ B'' := \frac{(S^2)^d_{23}}{(S^2)^d}$$

Subadditivity for d amounts to checking whether

$$\sqrt{A+B} \overset{?}{\leq} \sqrt{A'+B'} + \sqrt{A''+B''} \tag{11}$$

which holds by Lemma 5 because, by the induction hypothesis for $d-1$, we have $\sqrt{A} \leq \sqrt{A'} + \sqrt{A''}$ and, by the proof for $d = 1$, we have $\sqrt{B} \leq \sqrt{B'} + \sqrt{B''}$. □

Proof (Theorem 2): Non-negativity, reflexivity and symmetry are proven in a way analogous as in Theorem 1. As to subaddivity, we just need to prove the case $d = 1$, that is, the inequality analogous to Expression (9) for numerical variances. The proof for general d is the same as in Theorem 1. For $d = 1$, we have

$$S^2_{13} = \frac{(x_1 - x_3)^2}{2};\ S^2_{12} = \frac{(x_1 - x_2)^2}{2};\ S^2_{23} = \frac{(x_2 - x_3)^2}{2}$$

Therefore, Expression (9) obviously holds with equality in the case of numerical variances because

$$\sqrt{S^2_{13}} = \frac{x_1 - x_3}{\sqrt{2}} = \frac{(x_1 - x_2) + (x_2 - x_3)}{\sqrt{2}} = \sqrt{S^2_{12}} + \sqrt{S^2_{23}}$$

□

Acknowledgments and disclaimer

Thanks go to Klara Stokes for useful help. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

[1] J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In C. C. Aggarwal and P. Yu, editors, *Privacy-Preserving Data Mining: Models and Algorithms*, volume 34 of *Advances in Database Systems*, pages 53–80. New York: Springer, 2008.

[2] J. Domingo-Ferrer and A. Solanas. A measure of nominal variance for hierarchical nominal attributes. *Information Sciences*, 178(24):4644–4655. 2008. Erratum in *Information Sciences*, 179(20):3732, 2009.

[3] G. T. Duncan, M. Elliot and J.-J. Salazar-González. *Statistical Confidentiality: Principles and Practice*, New York: Springer, 2011.

[4] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf. *Handbook on Statistical Disclosure Control (version 1.2).* ESSNET SDC Project, 2010. `http://neon.vb.cbs.nl/casc`

[5] R. Lenz. *Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung*, Statistik und Wissenschaft, vol. 18, Wiesbaden: Statistisches Bundesamt, 2010.

[6] K. B. Reid. Centrality measures in trees. In *Advances in Interdisciplinary Applied Discrete Mathematics* (editors H. Kaul and H. M. Mulder), pp. 167-197, World Scientific eBook, 2010.

[7] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[8] L. Willenborg and T. DeWaal. *Elements of Statistical Disclosure Control.* New York: Springer, 2001.